Large language models encode clinical knowledge

Year: 2022 | Citations: 3213 | Authors: K. Singhal, Shekoofeh Azizi, T. Tu

Abstract

Med-PaLM, a state-of-the-art large language model for medicine, is introduced and evaluated across several medical question answering tasks, demonstrating the promise of these models in this domain. Large language models (LLMs) have demonstrated impressive capabilities, but the bar for clinical applications is high. Attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks. Here, to address these limitations, we present MultiMedQA, a benchmark combining six existing medical question answering datasets spanning professional medicine, research and consumer queries and a new dataset of medical questions searched online, HealthSearchQA. We propose a human evaluation framework for model answers along multiple axes including factuality, comprehension, reasoning, possible harm and bias. In addition, we evaluate Pathways Language Model^ 1 (PaLM, a 540-billion parameter LLM) and its instruction-tuned variant, Flan-PaLM^ 2 on MultiMedQA. Using a combination of prompting strategies, Flan-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-choice dataset (MedQA^ 3 , MedMCQA^ 4 , PubMedQA^ 5 and Measuring Massive Multitask Language Understanding (MMLU) clinical topics^ 6), including 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%. However, human evaluation reveals key gaps. To resolve this, we introduce instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars. The resulting model, Med-PaLM, performs encouragingly, but remains inferior to clinicians. We show that comprehension, knowledge recall and reasoning improve with model scale and instruction prompt tuning, suggesting the potential utility of LLMs in medicine. Our human evaluations reveal limitations of today’s models, reinforcing the importance of both evaluation frameworks and method development in creating safe, helpful LLMs for clinical applications.